Design and Development for enhancing Secure Big Data Environment With Different Protocols

Submitted to

Swami Ramanand Teerth Marathwada University, Nanded.

For the degree of

Doctor of Philosophy in computer science

Under

Faculty of Computer studies

By

Galal Adel Abdo Mohammed AL-Rummana

Under the guidance of

Dr. G. N. Shinde
Pro-Vice-Chancellor,
S.R.T.M.U,
Nanded- 431606.

INTRODUCTION

Background

Big Data is an abstract concept. As a part of masses of data, it has also some other features which determine the difference between itself and "massive data" or "very big data".

At present, the importance of big data has been generally recognized, but people still have different opinions on its definition. Also, different concerns such as scientific and technological enterprises, research scholars, data analysts, and technical practitioners have different definitions of big data [1].

The following definitions can provide a better understanding of big data:

In 2010, Apache Hadoop defined big data as "data sets which could not be captured, managed, and processed by general computers within an acceptable scope" [2]. In 2011, an IDC report defined big data as "big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis" [3].

Big data can be defined as a collection of data sets, which is very large in size and complex as well, Generally, size of the data is Petabyte and Exabyte. It exceeds the processing capability of conventional data management systems and software techniques. However, big data come with big values. Traditional database systems are not able to capture, store and analyze this large amount of data. As the internet is developing, amount of big data continue to develop. Big data analytics provide new ways for businesses and government to analyze unstructured data. Nowadays, Big data is one of the most common topics in IT industry. It is going to play an important role in the future. Big data changes the way that data is managed and used. Some of the applications are commonly used like, in healthcare, traffic management, banking, retail, education and so on. Organizations have become more flexible and more open. However, new types of data will give new challenges as well [4].

Characteristics of big data

Big Data has specific characteristics that affect information security: variety, volume, velocity, value, variability, and veracity as shown in fig (1) [5].

These characteristics can be presented as follows:

- Volume:

  Big data implies an enormous amount of data generated by machines, networks, and systems like social media etc. Analyzing such large amount of data is not an easy task.

- Variety:

  The variety characteristic refers to various types of structured and unstructured data originating from miscellaneous sources. Traditionally, data was stored structurally in database files or excel sheets, but today data comes in the form of images, audio files, videos, photos etc. This variety of unstructured data causes several problems in storing and analyzing data.

- Velocity:

  Velocity refers to the enormous rate at which the data is produced. This flow is massive and continuous. Analyzing real-time data at such a great speed is again a challenge to big data systems.

- Veracity:

  The data contains a lot of biases, noise, and abnormalities and thus, has to be cleaned before processing. Veracity refers to the resistance provided by big data to keep the "dirty data" from accumulating.

- Volatility:

  Validity refers to the point after which the data becomes futile for further analysis and hence, must be truncated.

- Validity:

  Validity tackles the issue of data trustworthiness and data precision. While analyzing the colossal amount of data, it is of importance that the data should be correct and accurate [6].

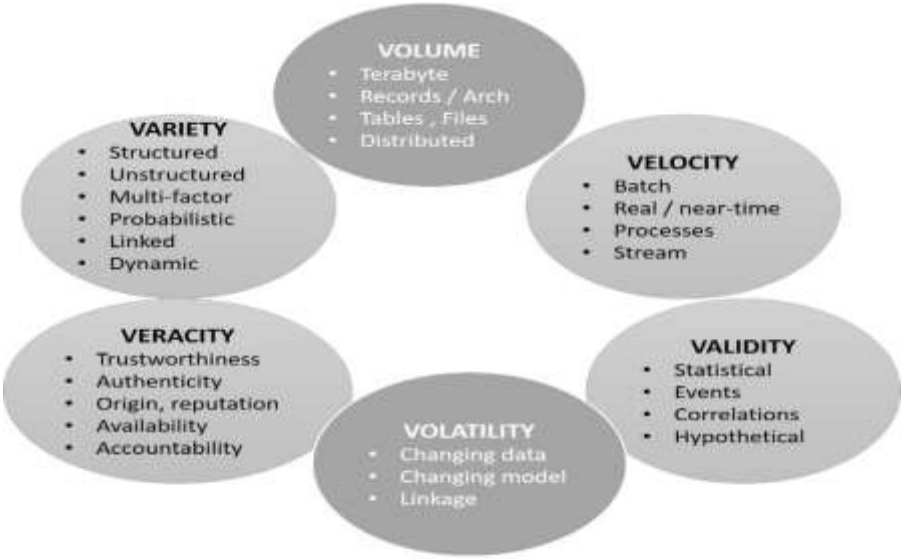

Fig (1) : The characteristics of big data

IMPORTANCE OF SECURITY IN BIG DATA

Given the volume and complexity involved with Big Data, providing security becomes a complicated task. Security is one of the most important parameters to consider in Big Data because if the data is intercepted by a rival business firm, it can gain insights and release policies which will combat the progress of the current firm [7]. There is three main security issue related to Big Data:

- Data Confidentiality

Data confidentiality is defined as the protection of data against unauthorized access. It refers also to distorted data from theft [8]. In other words, data confidentiality aims to implement access control mechanisms in order to authorize the users. Big Data invites data from a variety of sources and each source has its own unique access control policy. Merging a large number of access control policies under one common policy is a complicated task. Furthermore, to combat the velocity characteristic, Big Data authorization should be done automatically rather than manually (digital signatures).

- Data Integrity

Data trust worthiness should ensure that data is always consistent throughout the big data lifecycle [9]. Implementing Data provenance and data correlation techniques are challenging in big data [10].

- Data Privacy

Data privacy aims to protect the intruders from identifying information from the data base. With the advent of social media, the issue of Data Privacy has become momentous as Big data research is being carried out to create and analyze out profiles and if the gleaned information is not protected, it can be used for nefarious purposes [11].

LITERATURE SURVEY:-

The data security and privacy have become a critical issue in the hold of big data.

Many researchers have been attracted with securing data in transit, in storage, and in use.

In addition, how to allow on authenticated and authorized users to access and share data in a proper way in the term of big data is also a critical problem which attracts a researchers attentions. Nowadays, cloud computing is used as storage resources for big data. Both of them is considered as the backbone to the modern data storage and access mechanism.

The adoption of cloud technology for storing and managing big data and their applications adds yet another dimension to the problem of data security and privacy for big data[12].

Here, we provide an important review for existing security techniques and algorithm that approaching big data security

In 2016, Gai et al [13]. They have focused on the problem of the cloud data storage and aimed to provide an approach that could avoid the cloud operators reaching user' sensitive data. They proposed a novel approach entitled as Security-Aware Efficient Distributed Storage (SA-EDS) model. In their model, they used their proposed algorithms, including Alternative Data Dis- contribution (AD2), Secure Efficient Data Distributions (SED2) and Efficient Data Conflation (EDCon) algorithms. Their experimental evaluations had proved that their proposed scheme could effectively defend major threats from cloud-side. The computation time was shorter than current active approaches. Future work would address securing data duplications in order to increase the level of data availability since any of data center's down will cause the failure of data retrievals.

In other words, their main contributions are

- They propose a novel cryptography approach for delivering mass distributed storage by users' original data cannot be directly reached by cloud operators. The proposed method is an effectual cryptography means for defending malicious activities occurred on the cloud server.
- They proposed an efficient data split mechanism that does not produce big overheads, as well as ensures data retrievability.

The graphical representation of their model can be shown in figure (2)

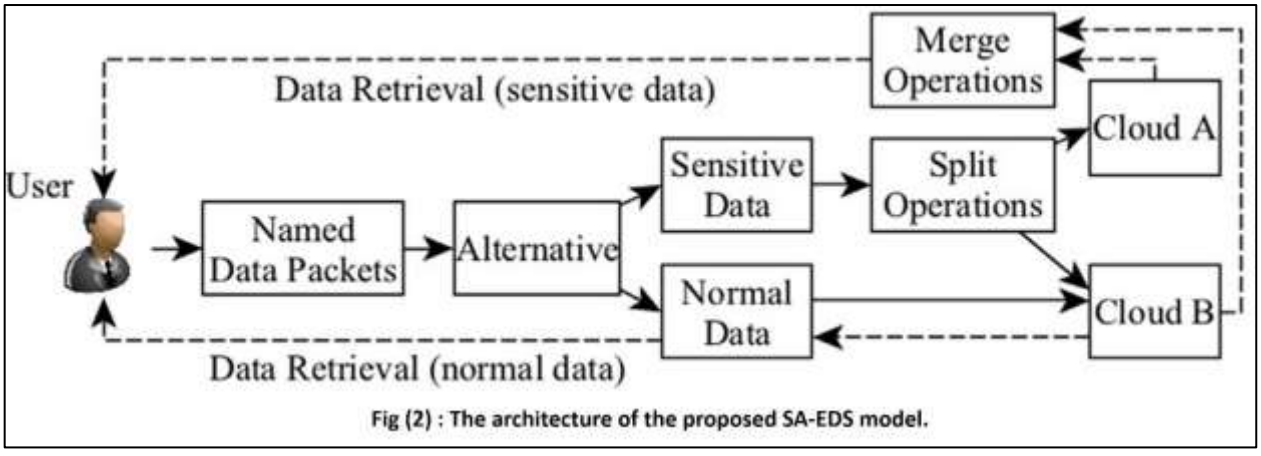

Fig (2) : The architecture of the proposed SA-EDS model.

The Big Data in the cloud is the most important research problem, in which researchers are trying to find the solution for it. One of the issues is to give a perfect security for big data in cloud computing, so that big data could be handled in the recent systems and managed with the cloud computing SreenivasaB.L et al [14].

Have studied and compared different encryption and decryption algorithm. Their comparison based on confidentiality, integrity, and availability of encryption and decryption security algorithm for big data in cloud computing.

In 2015, Hamlin et al [15]. Have presented a wide variety of cryptographic techniques for securing data in transit, in storage, and in use. They believe that, as security becomes a critical requirement for sensitive big data processing, these techniques will become an integral part of the big data ecosystem. They hope that the exposition the mentioned techniques will raise awareness of the latest types of tools and protections available for securing big data. They believe in better understanding and closer collaboration between the data science and cryptography communities will be critical to enable the future of big data processing.

Deng et al [16]. Proposed a Multi-user searchable encryption (MSE) which allows a user to encrypt his/her files in such a way that/these files can be searched by other authorized users.

Zhang et al [17]. Addresses the problem of popularity privacy leakage in social networks and develops a utility-based popularity anonymization (UPA) scheme to protect the privacy of users authentic popularity in online social networks without decreasing their utility. Their scheme utilizes the hierarchical authorization and capability delegation (HACD) model and proxy re-encryption with keyword search techniques to achieve k-anonymity of popularity. Extensive experiments on real-world online social networks confirm the efficiency of their scheme.

Regarding access control in big data, Access control is among the most effective solutions for full-fledged network security control. Conventional access control models such as role-based access control (RBAC), mandatory access control (MAC), and discretionary access control (DAC) are not applicable to support the access control in a untrusted domain like outsourcing environment. Generally, attribute-based access control (ABAC) is considered to provide flexibility and lightweight key management in data outsourcing environment. This is because the access control is based on the qualified set of attributes held by users and it offers the encryption feature in the model itself.

Data access control for big data should not only support the security but it should provide a reasonable performance for the retrieval and processing of encrypted data. This is because a vast amount of new data volume may be loaded and queried anytime. Therefore, general

attribute-based access control might not satisfy the performance requirement especially when there are a lot of attributes contained in the policy for encryption [18].

Traditional security perimeter such as a firewall is not sufficient to protect sensitive big data clusters since the data clusters may come from multiple sources and they may be distributed to several servers. Also, this cannot prevent malicious insider. Instead, the authorization enforcement such as attribute-based authorization according to the access control polity is an effective way to control the proper access. By utilizing cloud infrastructure, sensitive data needs to be encrypted for the privacy and security.

Fugkeaw and Sato [18]. Proposed a novel access control model combining Role-based Access Control (RBAC) model, symmetric encryption, and ciphertext attribute-based encryption (CP-ABE) to support fine-grained access control for big data outsourced in cloud storage systems. They also demonstrate the efficiency and performance of their proposed scheme through the implementation.

In [19]. The authors proposed access control model for big data based on Content-based Access Control (CBAC) model. Users can access the data based on the content rule defined in the CBAC policy. Also, the content similarity is also used to compare with a preset threshold, and the user is granted access to all the records that pass the similarity check. However, this approach does not assume the data to be outsourced in other domains. Security features are not discussed.

Consequently, Access Control systems are among the most critical of network security components. A system's privacy and security controls are more likely to be compromised due to the misconfiguration of access control policies rather than the failure of cryptographic primitives or protocols. This problem becomes increasingly severe as software systems become more and more complex, such as big data processing systems, which are deployed to manage a large amount of sensitive information and resources organized into a sophisticated big data processing cluster. Basically, big data access control requires the collaboration among cooperating processing domains to be protected as computing environments that consist of computing units under distributed access control managements.

Tools and techniques for big data access control should protect big data processes and data, ensuring that security policies are enforced in a cost-effective and timely manner. Currently, only a

few approaches that address the unique architecture of distributed computing can meet the security requirements of big data access control. Traditional approaches that rely on perimeter security are unable to adequately secure a big data cluster. For example, firewalls attempt to map IP to actual Active Directory credentials, but this is problematic in the big data cluster because it requires specific network design (i.e., no Network Address Translation (NAT) from internal corporate subnets). Even with special network configuration, a firewall can only restrict access on an IP/port basis and knows nothing of the architecture of the big data cluster. As a result, security administrators have to segregate sensitive data on separate servers in order to control access. It would require the creation of a second big data cluster to contain sensitive data, and even then would only provide two levels of security for the data. Even without those drawbacks, perimeter security solutions represent a single layer of defense around a soft interior [20].

Security professionals apply most controls at the very edges of the network. However, if attackers penetrate the big data cluster, they will have full and unrestricted access to the big data. Thus, many organizations are being required to enforce access control and privacy restrictions on big data to meet regulatory requirements.

Regarding authentication in big data, securing Big Data has become an important aspect as it can lead to disastrous results if intercepted by intruders. They have described the management of authentication and access control by providing an overview of the existing protocols. Further, they propose a modified protocol which amalgamates the authentication and authorization processes there by speeding up the entire procedure. Authentication and authorization are done in two separate ways in most systems thus requiring extra time and effort. To improve the efficiency of the system, in this work they have tried to coalesce the two pillars of security into a single one. Currently, the most widely used protocol for authentication is Secure Remote Password Protocol. They have proposed a model to provide dynamic attribute based access control by adopting the research that is done in a static role-based model and integrating it with the SRP Protocol. They have also elaborated the attacks that are repelled by the LV-SRP protocol by mathematically proving the repulsion. The mentioned model it is flexible to fit different attribute requirements as per the need of the organization. It strives to provide a stronger and complex access control mechanism due to the use cases that are usually enabled by them. It also makes endeavors to increase the speed at which the amalgamated process is performed in comparison to the traditional methods, Jhaveri et al [6]

AIMS:

Our aim is to design and development for enhancing secure big data environment with different protocols such algorithms to secure data in the big data environment. Developing a security system can be done through aspects of security such as:

- Access control (In order to provide an effective control to access data in big data environment with high performance and low computation.
- Data protection (In order to improve the exist encryption algorithms to be proper for big data environment.
- User's authentication (In order to protect data in big data environment from unauthorized users.

OBJECTIVES OF THE STUDY:

1. To provide overview of security in big data environment with different protocols based on related research works done in this area.
2. To investigate the existing works in security of big data environment with different protocols regarding to data protection, access control, and authentication.
3. To design and development security framework for big data environment with different protocols.
4. To implement the development framework using simulation tools.
5. To evaluate the performance and efficiency of the proposed protocol by comparing with existing protocol.
6. To give conclusion and recommendations for future research.

METHODOLOGY:

- Analyzing and evaluating the different existing techniques and algorithms that are related to big data security.
- Using such big data tools, which can support the security of big data environment.
- Using any simulator in order to build a new security system, which includes encryption, access control, and authentication algorithm to improve the security in the big data environment.
- Using a proper simulator for testing the validating the algorithms of the proposed security system.

PROPOSED WORK PLAN

Year-wise plan of work and targets to be achieved

Year 1:

- Literature Survey.
- Theoretical analysis of the existing security system applied to improve the security in big data environments

Year 2:

- Design and implementation a security system including data protection, access control, and authentication for big data environment.
- Performance evaluation of experimental results in terms of practical use.

Year 3:

- Result Analysis and Interpretation of the experimental result.
- Preparation of final synopsis and thesis.
- Submission of synopsis and thesis.

Mr. Galal Adel Abdo AL-Rummana **Dr.G.N.SHINDE**

(RESEARCH SCHOLLAR) **(RESEARCH GUIDE)**

REFERENCES:

[1] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171-209.

[2] Manyika J, McKinsey Global Institute, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute

[3] Gantz J, Reinsel D (2011) Extracting value from chaos. IDC view, pp 1–12

[4] Tarekegn, G. B., & Munaye, Y. Y. (2016). Big Data: Security Issues, Challenges, and Future Scope. International Journal of Computer Engineering & Technology (IJCET), 7(0976-6367), 12-24.

[5] Moura, J., & Serrão, C. (2016). Security and privacy issues of big data. arXiv preprint arXiv:1601.06206.

[6] Jhaveri, M., Jhaveri, D., & Shekokar, N. (2015). Big Data Authentication and Authorization using SRP Protocol. International Journal of Computer Applications, 130(1), 26-29.

[7] Akash Gupta, Alok Shukla, S.Venkatesanj, "Big Data: Cryptographically Enforced Access Control and Secure Communication" Proceedings of 6th IRF International Conference, Chennai, India, 10th May 2014, ISBN: 978-93-84209-16-2I.

[8] L. Xu, D. Sun, and D. Liu, "Study on methods for data confidentiality and data integrity in a relational database," 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT '10), vol. 1, pp. 292–295, July 2010.

[9] Akhil Mittal, "Trustworthiness of Big Data," International Journal of Computer Applications (0975 – 8887), Volume 80 – No.9, October 2013.

[10] Boris Glavic,"Big Data Provenance: Challenges and Implications for Benchmarking", Specifying Big Data Benchmarks, Volume 8163, Springer Berlin Heidelberg, 2014.

[11] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt, "Big Data Privacy Issues in Public Social Media", 6th IEEE International Conference on Digital Ecosystems Technologies (DEST), June 2012.

[12] Elisa Bertino, "Editorial: Introduction to Data Security and Privacy", *Data Science and Engineering*, pp. , 2016, ISSN 2364-1185.

[13] Li, Y., Gai, K., Qiu, L., Qiu, M., & Zhao, H. (2017). Intelligent cryptography approach for secure distributed big data storage in cloud computing. Information Sciences, 387, 103-115.

[14] SreenivasaB.L, Manish Kumar, Mohammed Nueed Shaikh, S. Sathyanarayana " A STUDY ON ENCRYPTION DECRYPTION ALGORITHM FOR BIG-DATA ANALYTICS IN CLOUD", International Journal of Latest Trends in Engineering and Technology Special Issue SACAIM 2016, pp. 323-329

[15] Hamlin, A., Schear, N., Shen, E., Varia, M., Yakoubov, S., & Yerukhimovich, A. (2016). Cryptography for big data security. Big Data: Storage, Sharing, and Security (3S), 241-288.

[16] Z. Deng, K. Li, K. Li, J. Zhou, A multi-user searchable encryption scheme with keyword authorization in a cloud storage, Future Gener. Comput. Syst. 72 (2017) 208–218.

[17] S. Zhang, Q. Liu, Y. Lin, Anonymizing popularity in an online social network with full utility, Future Gener. Comput. Syst. 72 (2017) 227–238.

[18] Fugkeaw, S., & Sato, H. (2015, November). Privacy-preserving access control model for big data cloud. In Computer Science and Engineering Conference (ICSEC), 2015 International (pp. 1-6). IEEE.

[19] Wenrong Zeng, Yuhao Yang, Bo Luo, Access Control for Big Data using Data Content, IEEE International Conference on Big Data, 2013

[20] Hu, V. C., Grance, T., Ferraiolo, D. F., & Kuhn, D. R. (2014, October). An access control scheme for big data processing. In Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2014 International Conference on (pp. 1-7). IEEE.